BA350 Econometrics Project : Regression analysis for the "Weather in Szeged 2006-2016" dataset

Ahmed Mnaouer

January 30, 2025

Abstract

This project investigates the relationship between humidity and various environmental factors using regression analysis. The dataset includes weather-related variables such as humidity, wind speed, pressure, visibility, and wind bearing. Exploratory data analysis (EDA) revealed strong correlations between humidity and some of these factors. Using various techniques, a potentially best regression model could be found and used to predict humidity (the dependent variable) for given levels of the chosen independent variables. Key insights include the identification of significant predictors and the validation of model assumptions. The findings offer actionable insights for weather prediction and analysis.

1 Introduction

Understanding the factors influencing humidity is crucial for weather prediction and climate studies. This project aims to analyze the relationship between humidity and environmental variables such as temperature, wind speed, and pressure. Using regression analysis, we identify key predictors and evaluate model performance. The dataset used in this study contains historical weather data, which was cleaned and preprocessed to ensure quality. The results provide valuable insights into the dynamics of humidity variation.

2 Methodology

2.1 Data Description

The dataset consists of historical weather data, including variables such as temperature, humidity, wind speed, and pressure. Data cleaning involved handling missing values, removing duplicates, and handling data type consistency. Feature engineering included extracting temporal features (e.g., hour, month) for further analysis. The complete code for data preprocessing, exploratory data analysis, and regression modeling is available in the link

2.2 Analytical Methods

In this project, we mainly used and evaluated Linear Regression models. The models were evaluated using metrics such as R-squared and root mean squared error (RMSE). Residual analysis, normality, autocorrelation and homoscedasticity tests were conducted to validate model assumptions. Linearity assumption was also verified in order to have the whole LINE assumptions validated for the final chosen model.

3 Analysis and Results

3.1 Exploratory Data Analysis (EDA)

Our EDA in this project consisted mainly of visualizing histograms, scatter plots, density plots, pair plots, and box plots for numerical columns (the ones to be used in the regression analysis) and the correlation

matrix. Time-series analysis was also conducted by plotting time series for numerical columns. Some examples of the plots can be shown in the figures below.

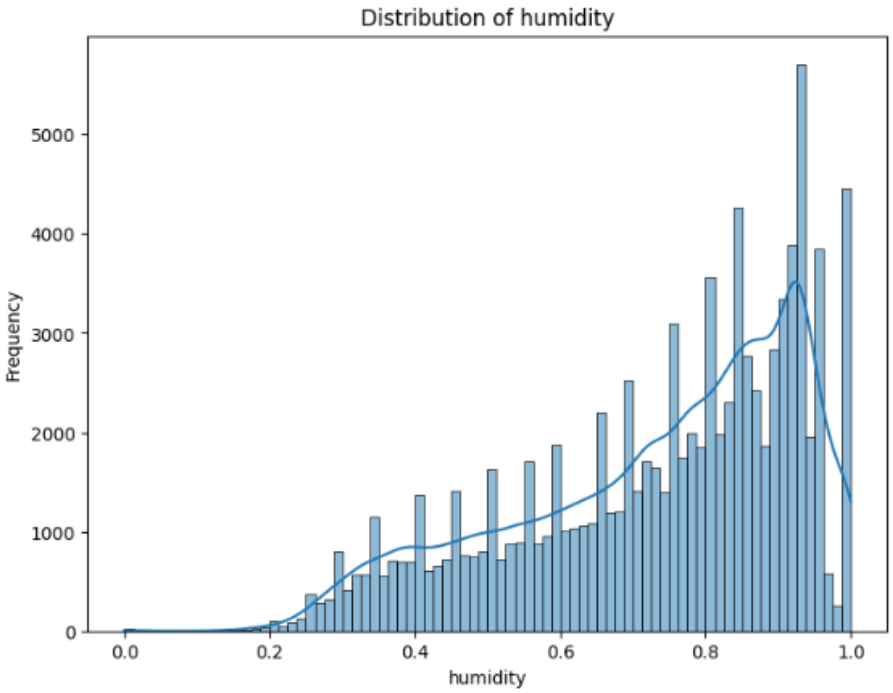

Figure 1: Distribution of Humidity

Figure 1 shows that humidity is highly skewed to the left.

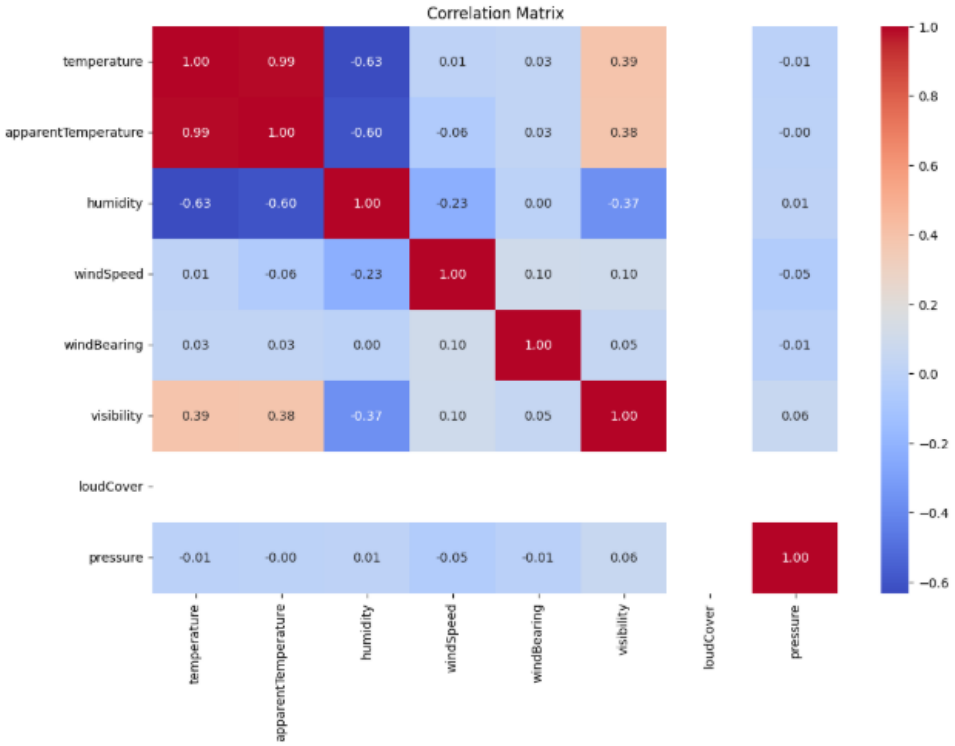

Figure 2: Correlation Matrix

Figure 2 shows the correlation matrix which indicates that humidity has a moderate to strong correlation with Temperature and Apparent Temperature, which are highly correlated to each other.

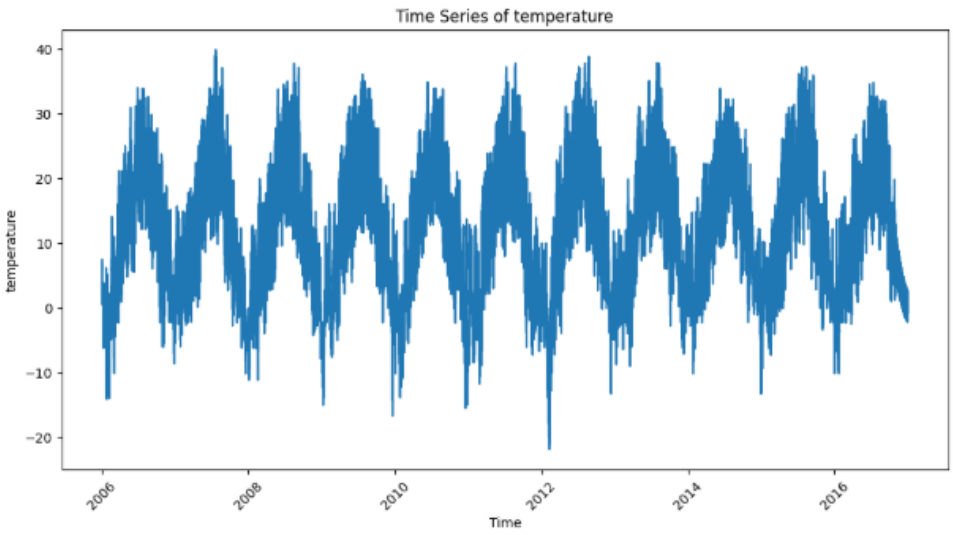

Figure 3: Distribution of Temperature

Figure 3 represents the time series for temperature which shows that it peaks at the half of each year, which is the summer period from each year.

Finally, Outlier detection using Z score was also performed and it showed 2538 outliers detected using Z score. Finally, we performed categorical column analysis and visualized pivot tables.

3.2 Regression Analysis

Before preceding to testing our different models, we tried to determine what can our dependent variable be from the list of available variables. The correlation matrix allowed to us to assess 3 potential variables : temperature, humidity and visibility.

Using the Features Importances graph for each variable, Humidity seemed to have the best set of possible regressors.

The following steps were then further conducted in order to ensure a good fitting model :

- Checking for multicollinearity calculating VIFs : loudCover, temperature, pressure and visibility were all eliminated due to their respective VIFs exceeding 5.

- The final list of independent variables to work with contains the followings : apparentTemperature, windSpeed, windBearing, hour, month.

- Multiple linear regression models were further assessed using the adjusted R-squared criterion with the help of the Lasso Regularization Technique.

Final Result : The best model in terms of adjusted R squared is the multiple regression model including the features Apparent Temperature, Wind Speed, Wind Bearing, Hour and Month.

Using the Ordinary Least Squares Method, we could determine the coefficients of our multiple linear regression equation :

Variable	Coefficient	p-value
Intercept	0.9023	0.000
Apparent Temperature	-0.0115	0.000
Wind Speed	-0.0071	0.000
Wind Bearing	8.581e-05	0.000
Hour	-0.0027	0.000
Month	0.0076	0.000

Table 1: Regression Results

Variables in the following Figure 4 represent:

- Y : humidity

- X1 : apparentTemperature
- X2 : windSpeed
- X3 : windBearing
- X4 : hour
- X5 : month

Final Equation : Y = 0.9023 -0.0115X1 - 0.0071X2 + 8.581e-05X3 - 0.0027X4 + 0.0076X5

Figure 4: Multiple Linear Regression Equation

The model accounts for 46 percent of the variation in humidity, as indicated by the adjusted R-squared in our analysis. Also, The fact that the R-squared and adjusted R-squared values are equal is a positive sign, suggesting no evidence of overfitting. This implies that all features in the model are relevant and contribute to explaining the dependent variable. Additionally, the small p-value and significant F-statistic confirm the model's overall significance.

Remark : A better performing polynomial model with 5 degrees using the Cross-Validation was found, explaining up to 67 percent of the variation in humidity using the specified features. This is an improvement over the previous multiple linear regression model, indicating that a non-linear relationship might be a better fit for the data.

However, Due to the complexity and difficulties in interpreting higher-degree polynomial regression, the model was simplified back to a linear regression (degree = 1). This simplification facilitates a clearer understanding and analysis of the relationships between the input variables and the target. Additionally, it helps mitigate risks of overfitting and reduces computational inefficiencies.

3.3 Residual Analysis and LINE Assumptions

- Linearity :

The residual plot (Figure 5) shows that the linearity assumption is not validated which is expected given that our best model is a polynomial model with 5 degrees, which indicates a non-linear pattern.

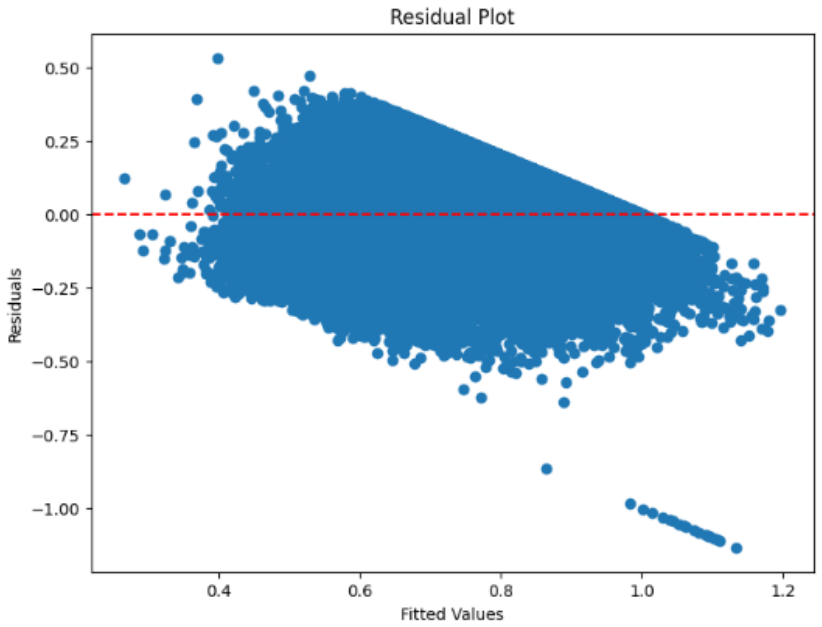

Figure 5: Residual Plot

- Independence of errors :

The computed Durbin Watson statistic is equal to 2, which indicates the absence of autocorrelation in the residuals. The Durbin Watson test in this case is useful since we have time series data.

- Normality :

The Residuals Histogram in the following Figure 6 suggests that the distribution of residuals is skewed to the left, resulting in a non-normal distribution within the residuals. Additionally, The computed Shapiro-Wilk p value suggests that the data being tested is significantly different from a normal distribution. Therefore, the normality assumption can not be validated.

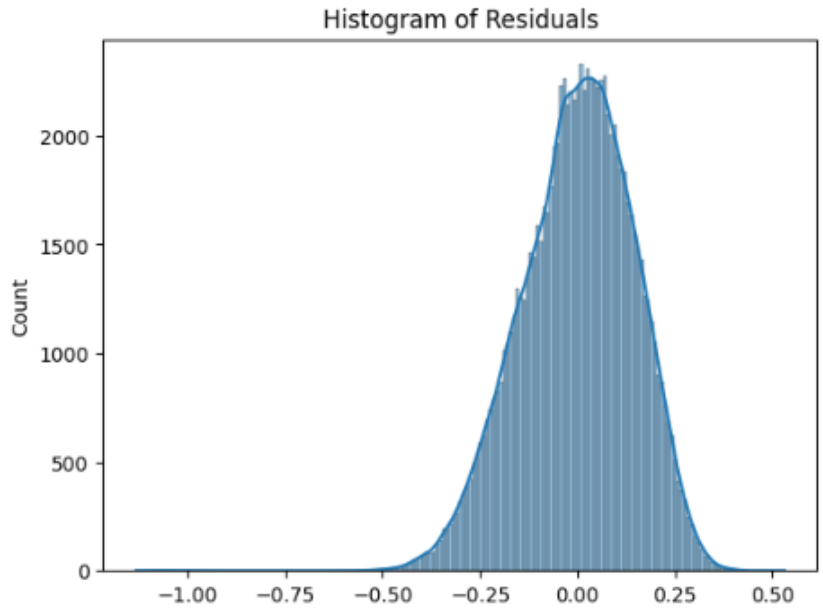

Figure 6: Residuals Histogram

- Homoscedasticity :

The computed Lagrange Multiplier statistic likely indicates strong evidence of heteroscedasticity in the regression analysis.

4 Discussion

All five predictor variables demonstrated statistical significance (p ¡ 0.05), indicating that each variable contributes meaningful information to the prediction of humidity. While diagnostic plots revealed some deviations from classical linear regression assumptions, particularly in terms of homoscedasticity and linearity, this is not uncommon when modeling environmental variables such as humidity, which often exhibit complex, non-linear relationships.

The model achieved meaningful predictive capability, with residuals largely falling within ±0.5 units of actual humidity values. This practical accuracy, combined with the statistical significance of all predictors, suggests the model captures important relationships in the data, even if these relationships might not be strictly linear.

As explained in the 3.2 section above, a better polynomial model was found, which can be the cure to these violations. However, the current model provides valuable insights into the relationships between the predictor variables and humidity, while maintaining interpretability, a key consideration for practical applications.

5 Actionable Insights

Interpretation of the coefficients :

- **const (0.9023)**: This is the intercept. When all independent variables are zero, the predicted humidity is approximately 0.90.

- **apparentTemperature (-0.0115)**: For every unit increase in apparent temperature, humidity decreases by approximately 0.0115 units, holding other variables constant. This variable is statistically significant (low p-value).
- **windSpeed (-0.0071)**: For every unit increase in wind speed, humidity decreases by approximately 0.0071 units, holding other variables constant. This variable is statistically significant (low p-value).
- **windBearing (8.581e-05)**: For every unit increase in wind bearing, humidity increases by a very small amount (8.581e-05 units), holding other variables constant. Although statistically significant, the effect is likely negligible in practice.
- **hour (-0.0027)**: For every unit increase in hour, humidity decreases by approximately 0.0027 units, holding other variables constant. This variable is statistically significant (low p-value).
- **month (0.0076)**: For every unit increase in month, humidity increases by approximately 0.0076 units, holding other variables constant. This variable is statistically significant (low p-value).

Approximately 46 percent of the variance in the dependent variable (humidity) is explained by the model.

Using our multiple regression model :

- Humidity is predicted to be equal to 0.73 when the apparent temperature is 15 degrees Celsius, wind speed is 5 m/h, wind bearing is 180 degrees, at 10 am in the morning in June.

- It is also predicted to be equal to 0.57 when the apparent temperature is 25 degrees Celsius, wind speed is 15 m/h, wind bearing is 360 degrees, at 6 pm in October.

Beyond the statistical findings, these insights have practical implications: Weather forecasting models can integrate these relationships to improve short-term humidity predictions. Agricultural planning can benefit from understanding how temperature and wind speed affect humidity levels, helping optimize irrigation schedules. Similarly, urban planners can use this data to design efficient cooling and ventilation systems.

The Polynomial regression model with degree 5 provides a better fit compared to the simple linear regression model as indicated by higher R-squared and adjusted R-squared values. This suggests a non-linear relationship between the predictors and humidity.

6 Conclusion

This project aimed to explore the relationship between humidity and various environmental factors using regression analysis. Through a comprehensive methodology involving data cleaning, exploratory data analysis (EDA), and rigorous model evaluation, we identified key predictors of humidity, including apparent temperature, wind speed, wind bearing, hour, and month. The final multiple linear regression model, while simplified for interpretability, explained approximately 46 percent of the variance in humidity, with all predictor variables demonstrating statistical significance.

Despite deviations from some classical linear regression assumptions—such as linearity, normality, and homoscedasticity—the model provided meaningful insights into the complex relationships between environmental variables and humidity. The polynomial regression model, which explained 67 percent of the variance, further highlighted the non-linear nature of these relationships. However, for practical applications, the linear model was retained due to its simplicity and ease of interpretation.

The actionable insights derived from this analysis, such as the inverse relationship between apparent temperature and humidity, can be leveraged in various real-world applications, including weather forecasting, agricultural planning, and urban design. While the model has limitations, it serves as a robust foundation for further research, particularly in exploring advanced machine learning techniques or incorporating additional variables to improve predictive accuracy.

In conclusion, this study underscores the importance of regression analysis in understanding environmental dynamics and provides a framework for future investigations into humidity prediction and its implications for climate studies and practical applications.

References

- Python Software Foundation. (2023). Python Language Reference. Retrieved from https://www.python.org/

- Statsmodels Development Team. (2023). Statsmodels: Statistical modeling in Python. Retrieved from https://www.statsmodels.org/
- Scikit-learn Development Team. (2023). Scikit-learn: Machine learning in Python. Retrieved from https://scikit-learn.org/
- Kaggle. (2023). Weather in Szeged 2006-2016.
  Retrieved from : https://www.kaggle.com/datasets/budincsevity/szeged-weather/data
  CSV file : 'weatherHistory.csv'
- Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach* (7th ed.). Cengage Learning.
- Greene, W. H. (2018). *Econometric Analysis* (8th ed.). Pearson.
- Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (3rd ed.). Academic Press.
- McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts. Retrieved from https://otexts.com/fpp3/